Prediction-Driven Staffing for Emergency Departments: What to Predict and How to Predict

Lin (Franklin) Feng
Graduate School of Business
Stanford University
Stanford, CA 94305
franklin.feng@stanford.edu

Jing Dong
Graduate School of Business
Columbia University
New York, NY 10027
jing.dong@gsb.columbia.edu

Abstract

We conduct a comparative study of prediction-driven strategies for staffing hospital emergency departments (ED). We evaluate three approaches: (i) a *machine learning (ML)* approach that relies on census forecasts and applies a straightforward patient-to-nurse ratio to determine staffing; (ii) a *staffing-level informed machine learning (SIML)* approach that models the mapping from staffing levels to congestion outcomes and chooses the staffing plan that minimizes the associated cost; and (iii) a *queueing-informed (QI)* approach that leverages a calibrated queueing model to guide staffing decisions. We evaluate the three approaches using real ED arrival patterns. ML, which overlooks the endogeneity of queueing dynamics, can suffer from varying degrees of delayed feedback. SIML performs well when training and evaluation conditions align, but can be sensitive to distribution shifts. QI typically achieves the best results under correct model specification, though it is vulnerable to misspecification, for which we provide a diagnostic tool. Finally, we offer practical guidance to help hospitals select the most suitable approach given their data and modeling expertise.

1 Introduction

Emergency department (ED) crowding has become a widespread crisis, exacerbated by a national shortage of nurses [Hoot and Aronsky, 2008]. Chronic understaffing leaves EDs unable to meet surges in demand, leading to long waits and patients leaving without care. At the same time, hospitals face intense budgetary pressures that limit their ability to overstaff [Hodgson et al., 2024]. These challenges make it urgent to develop prediction-informed staffing algorithms that help hospitals allocate limited nursing resources more effectively, thereby reducing crowding and improving patient safety. In this paper, we examine prediction-driven strategies for ED nurse staffing, with the objective of minimizing total cost, defined as the sum of congestion-related waiting costs and staffing costs.

A central challenge in ED staffing is the feedback between staffing and congestion: staffing decisions influence not only the immediate system state but also future crowding, and approaches that ignore this dynamic often underperform. This challenge is further complicated by the fact that the ED is a highly non-stationary environment, with demand fluctuating across hours. At the same time, staffing levels must be fixed within each shift and scheduled in advance, restricting the ability to make real time adjustments. Against this backdrop, we examine three prediction-driven methods that capture the staffing-congestion interaction to varying degrees while also spanning the spectrum of implementation complexity faced by hospitals. I) Machine learning (ML): The simplest approach. In this setting, machine learning is used only for short-term forecasting of patient census, which is then converted into staffing through a fixed patient-to-nurse ratio. This design is intentionally simple,

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: MLxOR: Mathematical Foundations and Operational Integration of Machine Learning for Uncertainty-Aware Decision-Making.

reflecting common practice where hospitals combine forecast tools with ratio-based staffing rules, but it ignores how staffing itself shapes congestion. II) Staffing-informed machine learning (SIML): A more sophisticated approach that conditions predictions on candidate staffing levels, learning how different staffing choices affect congestion and selecting the plan that minimizes implied cost. This better accounts for feedback but requires richer training data and more careful model development. III) Queueing-informed (QI): The most structurally grounded but also the most demanding approach. It calibrates a queueing model to data, simulates patient flow under alternative staffing vectors, and selects the lowest-cost plan. When well specified, it fully internalizes staffing-congestion interactions, but calibration and validation require greater modeling expertise.

Our comparative analysis highlights implementation-relevant insights across a wide variety of scenarios, including settings with different levels of demand-prediction accuracy, varying baseline congestion, and alternative boarding conditions. Using real ED arrival patterns, we find that both SIML and QI substantially reduce total cost relative to ML in most cases. ML underperforms because it ignores the feedback between staffing and congestion, which manifests as delayed adjustments in system performance. SIML performs well when training and evaluation conditions are aligned but deteriorates under distributional shifts, such as when demand or staffing patterns differ from those represented in the training data. QI generally achieves the lowest costs when its queueing assumptions are valid, but its performance degrades under misspecification (e.g., alternative service-time distributions or omitted system features). To mitigate this risk, we develop a diagnostic test that compares simulated trajectories to observed data and signals when refinement is needed. Finally, we translate these findings into practical guidance, identifying which approach is most suitable given a hospital's data availability and modeling expertise.

This study connects to two main streams of literature. First, queueing models have long been used to study capacity sizing and staffing in service systems, including EDs. Foundational work traces back to Erlang's formulas and the square-root safety staffing principle [Erlang, 1917, Halfin and Whitt, 1981, Kolesar and Green, 1998, Borst et al., 2004]. In settings with time-varying demand and patient abandonment, analytical models have been developed to balance service quality with staffing cost [Garnett et al., 2002, Mandelbaum and Zeltyn, 2009, Gurvich et al., 2008, Green et al., 2007], and related prescriptions have been applied to ED operations [Green et al., 2006]. Second, there is a growing body of data-driven and machine learning research for forecasting and decision support in EDs. Studies develop statistical and learning-based methods for arrivals and census [Ang et al., 2016, Hu et al., 2021, Bacchi et al., 2020, Harrou et al., 2020] and link predictions to operational levers such as staffing and flow management [Xu and Chan, 2016, Wang et al., 2022]. A related stream integrates prediction with prescription by training models in ways that account for downstream decision quality [Oroojlooyjadid et al., 2020, Chen et al., 2023, Sir et al., 2017].

Our contribution is to bridge these streams through a single, implementable comparison of three approaches. The comparative study identifies when structural queueing information or staffing-informed learning provides advantages, quantifies sensitivity to misspecification and distribution shift, and provides practical guidance for method selection in hospitals.

2 Methods

We partition each day into N shifts (in our experiments $N = 6$), and select integer staffing levels $\mathbf{s} = (s_1, \ldots, s_N)$, fixed within each shift and determined at the start of the day. Total cost combines a service-performance component C_w, increasing with congestion through waiting and abandonment, and a staffing component $C_s(\mathbf{s})$, reflecting labor expenditure. Our baseline model of the ED is an $M_t/M/n+M$ queue, i.e., a multi-server queue with time-varying arrival rate and abandonment. The intra-day arrival pattern is represented by a fixed profile $\{\gamma_t\}_{t=0}^{23}$. Day-to-day variability is captured by a multiplicative scale factor ρ_D, so that the arrival rate on day D hour t is $\rho_D\gamma_t$. We assume γ_t is known, while ρ_D must be forecast with varying degrees of accuracy.

ML Approach The machine learning (ML) approach sets staffing based solely on forecasts of the ED census, without considering how staffing affects future patient flow. Let x_t denote the average census during hour t and a_t the arrival rate in hour t. At hour 0, the feature vector consists of the predicted arrival rate $\hat{a}_0$ and the most recent 48 hours of arrivals and census values: $(\hat{a}_0, (a_{-1}, x_{-1}), \ldots, (a_{-k}, x_{-48}))$. A feedforward neural network is trained to predict the one-step-ahead census x_0. Multi-step forecasts for the next 24 hours are then obtained autoregressively by

iteratively feeding predictions back as inputs. Staffing levels, s_{ML}, are obtained by dividing the predicted census by a fixed patient-to-nurse ratio.

SIML Approach The staffing-level informed machine learning (SIML) approach extends ML by incorporating candidate staffing levels into the prediction process. Let q_t denote the average queue length and a_t the arrival rate in hour t. At hour 0, the feature vector includes the predicted arrival rate $\hat{a}_0$, the candidate staffing level s_0, and the most recent 48 hours of arrivals, queue lengths, and staffing values: $(\hat{a}_0, s_0, (a_{-1}, q_{-1}, s_{-1}), \ldots, (a_{-k}, x_{-k}, s_{-k}))$, $k = 48$. The neural network is trained to predict the one-step-ahead queue length q_0, and forecasts for the next 24 hours are again generated autoregressively. The resulting queue length trajectory yields a proxy estimate of the service-performance cost C_w. The SIML staffing plan $\mathbf{s}_{\text{SIML}}$ is then chosen as the candidate vector that minimizes this proxy cost plus the staffing cost $C_s(\mathbf{s})$.

QI Approach The queueing-informed (QI) approach embeds an explicit structural model of ED operations (e.g., $M_t/M/n + M$ queue in the baseline scenario) into the staffing decision. Parameters such as service rates and abandonment are estimated from data, and each candidate staffing plan $\mathbf{s}$ is evaluated by simulating the induced queueing dynamics. The expected service-performance cost C_w is approximated by averaging across multiple simulation replications, which is then combined with the staffing cost $C_s(\mathbf{s})$ to yield an estimated total cost. The QI staffing plan $\mathbf{s}_{\text{QI}}$ is chosen as the vector that minimizes this estimated objective.

Evaluation. Performance is evaluated in a Monte Carlo framework spanning a wide range of operating regimes. Specifically, we compare short and long boarding scenarios, represented in the model by shorter versus longer service and abandonment times. Additional experiments vary the staffing cost, initial system load (starting census), and the accuracy of demand forecasts. To assess robustness, we further stress-test each method under scenarios with distributional shifts between training and deployment as well as model misspecification in the structural assumptions.

3 Results

Main experiment. In the main experiment, we assume that training and testing data are well aligned and that the queueing model used by QI is correctly specified. Under these benchmark conditions, QI generally achieves the lowest cost, with SIML performing closely behind and ML trailing. A notable exception occurs in long-boarding regimes, where congestion evolves so slowly that delayed feedback is less limiting. In this setting, census forecasts provide a stable proxy for future workload, allowing the ML approach to remain well aligned with realized demand. By contrast, SIML is disadvantaged because the slow dynamics obscure staffing-congestion effects, limiting the predictive model's ability to learn their impact and weakening its optimization.

We evaluate 14 scenarios that vary in initialization, staffing cost, boarding times, and prediction errors, we summarize the resulting performance gaps in Figure 1. Specifically, we compare $(\text{Cost}_{\text{SIML}} - \text{Cost}_{\text{QI}})/\text{Cost}_{\text{QI}}$ (SIML gap) and $(\text{Cost}_{\text{ML}} - \text{Cost}_{\text{QI}})/\text{Cost}_{\text{QI}}$ (ML gap) across these settings. In most cases, SIML closely tracks QI, with gaps below 2.5%. The main exceptions are scenarios B1, B2, and D1, where SIML is more vulnerable to distributional shifts. By contrast, the ML approach generally performs worse, with an average gap of 6.8% across all scenarios. Its gap narrows when boarding times are long, because delayed feedback is less limiting, and when staffing costs are low, because ML tends to overstaff and the penalty for doing so is smaller.

QI suffers from model misspecification risk To evaluate the robustness of QI to model misspecification, we consider two scenarios: (i) the structural form of the model is correct, but distributional assumptions are misspecified (i.e., exponential vs. lognormal/Erlang/Weibull service times); and (ii) the structural form itself is misspecified (e.g., omission of abandonment behavior). Distributional misspecification increases QI's cost by about 4-8% relative to correct specification, while structural misspecification raises costs by 4-5%. In most cases, QI still outperforms SIML, though the margin narrows. To help practitioners detect problematic misspecifications, we implement a trajectory-matching diagnostic that compares simulated queueing trajectories to historical ED data using dynamic time warping. The method produces a similarity score on a 0–100 scale: values above roughly 80 are associated with small performance losses ($\leq 5\%$), while lower scores reliably flag specifications that require refinement.

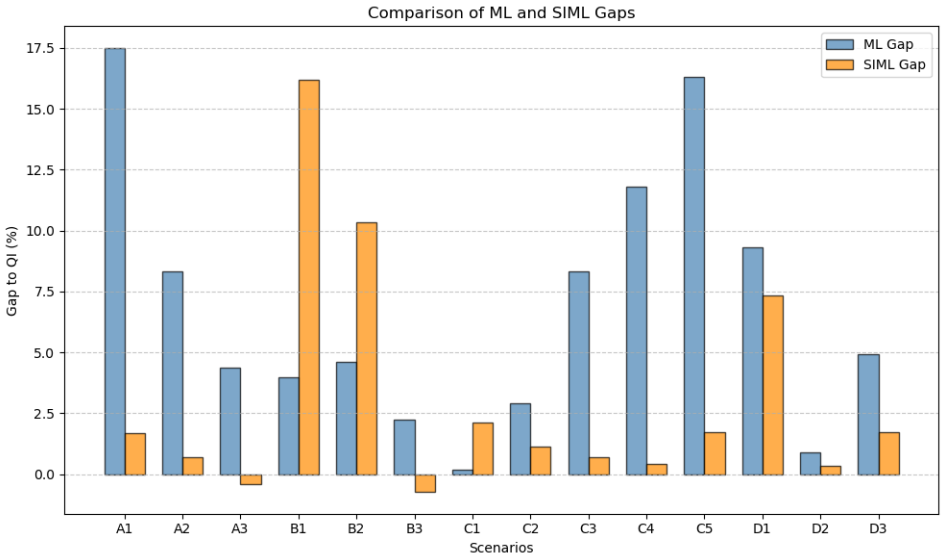

Figure 1: Gaps of mean cost between SIML/ML and QI across all scenarios: A1-A3 and B1-B3 correspond to short/long boarding time scenarios; C1-C5 correspond to different staffing costs; D1-D3 correspond to different levels of demand prediction error.

SIML suffers from distributional shift risk We evaluate SIML's robustness under three types of distributional shift between training and deployment. First, when training arrival rates cover a narrower range than those encountered at test time, the performance gap to QI grows modestly, from 7.5% at baseline to 8.6% under the widest expansion. Second, when testing arrivals are systematically higher than in training, performance deteriorates sharply: a 100% uplift in demand increases SIML's cost by 81% relative to its baseline and by nearly 50% relative to QI. Third, when training data are limited to extreme staffing regimes, SIML performs worst under highly overstaffed training ($\approx 20\%$ worse than baseline), as the absence of congestion reduces its ability to learn staffing-congestion interactions. Together, these experiments show that SIML's predictive mapping becomes unreliable when test conditions fall outside the training distribution, leading to degraded performance.

Executive Summary Taken together, our results provide clear guidance on how hospitals can select among the three prediction-driven staffing approaches. The ML approach, while the easiest to train and deploy, is generally outperformed by SIML and QI because it ignores the feedback between staffing and congestion. Its competitiveness improves only in slow-moving systems with long service or boarding times, where delayed feedback has less impact. The SIML approach performs better by conditioning predictions on staffing levels and capturing their effect on congestion. It works well when training and deployment conditions are aligned, but its accuracy deteriorates under distributional shifts, such as unexpected surges in demand or staffing patterns outside the training range. Because it depends on retraining and coverage of relevant operating conditions, SIML is data-intensive and best suited for hospitals with stable operations and strong predictive infrastructure. The QI approach generally achieves the lowest costs when its model is well specified and calibrated to data, as it fully internalizes the staffing-congestion interaction. However, it is vulnerable to misspecification and requires substantial queueing expertise to select, calibrate, and validate the underlying model. To mitigate this risk, we introduce a diagnostic tool that compares simulated trajectories to observed data and signals when refinement is needed.

In practice, hospitals should align method choice with their resources and capabilities. Institutions with limited data and a need for quick deployment may start with ML; those with rich historical data and the ability to retrain frequently may benefit from SIML; and those with the analytical capacity and queueing expertise can gain the most from QI.

References

Erjie Ang, Sara Kwasnick, Mohsen Bayati, Erica L Plambeck, and Michael Aratow. Accurate emergency department wait time prediction. *Manufacturing & Service Operations Management*, 18(1):141–156, 2016.

Stephen Bacchi, Samuel Gluck, Yiran Tan, Ivana Chim, Joy Cheng, Toby Gilbert, David K Menon, Jim Jannes, Timothy Kleinig, and Simon Koblar. Prediction of general medical admission length of stay with natural language processing and deep learning: a pilot study. *Internal and emergency medicine*, 15:989–995, 2020.

Sem Borst, Avi Mandelbaum, and Martin I Reiman. Dimensioning large call centers. *Operations research*, 52(1):17–34, 2004.

Xinyun Chen, Yunan Liu, and Guiyu Hong. Online learning and optimization for queues with unknown demand curve and service distribution. *arXiv preprint arXiv:2303.03399*, 2023.

Agner Krarup Erlang. Solution of some problems in the theory of probabilities of significance in automatic telephone exchanges. *Post Office Electrical Engineer's Journal*, 10:189–197, 1917.

Ofer Garnett, Avishai Mandelbaum, and Martin Reiman. Designing a call center with impatient customers. *Manufacturing & Service Operations Management*, 4(3):208–227, 2002.

Linda V Green, Joao Soares, James F Giglio, and Robert A Green. Using queueing theory to increase the effectiveness of emergency department provider staffing. *Academic Emergency Medicine*, 13 (1):61–68, 2006.

Linda V Green, Peter J Kolesar, and Ward Whitt. Coping with time-varying demand when setting staffing requirements for a service system. *Production and Operations Management*, 16(1):13–39, 2007.

Itay Gurvich, Mor Armony, and Avishai Mandelbaum. Service-level differentiation in call centers with fully flexible servers. *Management Science*, 54(2):279–294, 2008.

Shlomo Halfin and Ward Whitt. Heavy-traffic limits for queues with many exponential servers. *Operations research*, 29(3):567–588, 1981.

Fouzi Harrou, Abdelkader Dairi, Farid Kadri, and Ying Sun. Forecasting emergency department overcrowding: A deep learning framework. *Chaos, Solitons & Fractals*, 139:110247, 2020.

Nicole R Hodgson, Richard Kwun, Chad Gorbatkin, Jeanie Davies, Jonathan Fisher, and ACEP Emergency Medicine Practice Committee. Emergency department responses to nursing shortages. *International Journal of Emergency Medicine*, 17(1):51, 2024.

Nathan R Hoot and Dominik Aronsky. Systematic review of emergency department crowding: causes, effects, and solutions. *Annals of emergency medicine*, 52(2):126–136, 2008.

Yue Hu, Carri W Chan, and Jing Dong. Prediction-driven surge planning with application in the emergency department. *Submitted to Management Science*, 2021.

Peter J Kolesar and Linda V Green. Insights on service system design from a normal approximation to erlang's delay formula. *Production and Operations Management*, 7(3):282–293, 1998.

Avishai Mandelbaum and Sergey Zeltyn. Staffing many-server queues with impatient customers: Constraint satisfaction in call centers. *Operations research*, 57(5):1189–1205, 2009.

Afshin Oroojlooyjadid, Lawrence V Snyder, and Martin Takáč. Applying deep learning to the newsvendor problem. *IISE Transactions*, 52(4):444–463, 2020.

Mustafa Y Sir, David Nestler, Thomas Hellmich, Devashish Das, Michael J Laughlin Jr, Michon C Dohlman, and Kalyan Pasupathy. Optimization of multidisciplinary staffing improves patient experiences at the mayo clinic. *Interfaces*, 47(5):425–441, 2017.

Kanix Wang, Walid Hussain, John R Birge, Michael D Schreiber, and Daniel Adelman. A high-fidelity model to predict length of stay in the neonatal intensive care unit. *INFORMS journal on computing*, 34(1):183–195, 2022.

Kuang Xu and Carri W Chan. Using future information to reduce waiting times in the emergency department via diversion. *Manufacturing & Service Operations Management*, 18(3):314–331, 2016.